
Editorial

Hocine Cherifi*

Faculté des Sciences Mirande,
Université de Bourgogne,
LE2I UMR CNRS 6306,
BP 47870, Dijon, France
E-mail: hocine.cherifi@u-bourgogne.fr
*Corresponding author

Rushed Kanawati

Laboratoire d'informatique de Paris NORD (LIPN),
University Paris 13,
Sorbonne Paris Cité,
F-93430 Villetaneuse, France
E-mail: rushed.kanawati@lipn.univ-paris13.fr

Piet Kommers

Faculty of Behavioral Sciences,
University of Twente,
7500AE Enschede, The Netherlands
E-mail: kommers@edte.utwente.nl

Biographical notes: Hocine Cherifi received his MSc and PhD in Computer Science, both from the National Polytechnic Institute in Grenoble, France, in 1981 and 1984, respectively. He is presently a Professor in the Department of Computer Science, University of Burgundy in Dijon, France. His research activities are in the fields of computer vision, pattern recognition and complex networks.

Rushed Kanawati is an Associate Professor at the University of Paris 13, Sorbonne Cité. He received his MSc and PhD in Computer Science, both from the National Polytechnic Institute in Grenoble, France, in 1993 and 1997, respectively. His research topics cover machine learning, case-based reasoning and social network mining.

Piet Kommers is an Associate Professor at the University of Twente, The Netherlands. His specialty is social media for communication and organisation. As conference co-chair of the IADIS multi-conference, he initiated the conferences of web-based communities and social media, e-society, mobile learning and international higher education. He is a Professor at the UNESCO Institute for Eastern European Studies in Educational Technology and currently a Research Fellow at Curtin University in Perth, Australia.

This special issue is totally dedicated to the question how communities can be regarded as embedded structures in surrounding complex networks. Due to ‘Dunbar’s Number’ (1993), if you try to imagine how two random persons who do not know each other might relate to each other via indirect paths of relationships, you will soon find out that it opens a so-called combinatory explosion as already each of the persons has a circle of 400 to 3,000 persons typically they ‘know’ (Russell et al., 1987; Bialik, 2007). Within six steps deep each person on this globe ‘knows’ each other person (Milgram, 1967). However, to find the track via which the relationships develop is hard to find as it belongs to the family of NP complete problems (Sipser, 1997). One of the more prominent exemplars of NP-completeness is the subgraph isomorphism problem (Snijders et al., 2006.)

The relevance of graph computation for social network analysis has not been disputed. Still, it is worth to look for the most vital urgencies why it needs to be undertaken in the targeted readership of this journal just now. The first reason is that the study of social network analysis is needed in order to identify structural patterns, locate central (influential) network players and in order to detect how networks evolve along time. Exactly as we look to social networks underlying social media it is clear that dynamics is its main nature. We may speak about transient networks as its intermediate stages are even more important than its final outcome. Especially as the web of group affiliations expand in size, it is no longer possible to focus on the layout of sociograms; Its outcomes need to be formalised in terms of parameters that can be forwarded to social search engines like Wajam, Folkd, Slangwho, Sproose, Mahalo, Jumper 2.0, Qitera, Scour, Wink, Eurekster, Baynote, Delver, OneRiot, and SideStripe. Its main conceptual underpinning has already been articulated in the tradition of sociograms. It allowed group managers and for instance teachers to depict dependency relations and find out who are the leaders and who are the linking pins. Social network analysis a still growing area and is becoming epitomised as ‘network science’.

Web-based communities have been developing into three directions mainly. The first are those that ever more rely on societal-, ideational- and altruistic sentiments. Good examples are the communities for sharing the latest news and understanding for the sake of aggregating the best possible interpretation of the world at that very moment. There is a sense of journalism and also there is a large trend towards collectivism and constructivism. Reality can only be understood if contrasting ideas are allowed and even better, if oppositional ideas are provoked. The second direction of web-based communities is the functional one; one or several task domains are supported by its members in so-called ‘communities of practice’. Its functioning rests upon the experience that the transfer of expertise should not be trusted to the traditional top-down publishing and curricular mechanisms as they cost too much time before professional ‘best practices’ are consolidated. The community metaphor has proven to be more versatile and more to the point than the formalised ways of training and instruction for instance. The third direction is web-based communities to evolve as creational networks that target the restructuring of society because of global evolutions like after 9/11 and recently when the economic crisis started to develop. Topics with a similar magnitude are the shift from traditional- into problem-based education. But also the intrinsic trend nowadays to focus technology more and more on the aspect of sustainability like we see in the so-called ‘cradle-to-cradle’ engineering and design. A good example of how new network structures help to address new opportunities for social-/economic restructuring is the emergence of ‘creative industries’. It allows the art- and entertainment industry to get involved with genres in for instance gaming and simulations, that would not have been

possible before the notion of ‘network society’ was coined. In conclusion, this special issue of the *Journal of Web Based Communities* provides you with the graph theoretical underpinning of ‘network sciences’.

Business relations between companies, neural networks, metabolic networks, food webs, distribution networks such as the electric power grid, the internet, The World Wide Web, social networks, online chat are typical examples of complex systems. That is, a large number of individual elements interacting in order to achieve a global goal with just a local knowledge of the overall system and minimal communications with each other. Understanding the characteristics that hold together the parts of this various complex systems cannot rely on local knowledge on individuals but it must be based on the overall network effects caused by their interconnections. Indeed, all these systems can be adequately described using graph theory where individuals are the node (vertices) of the graphs and a link (edge) between two nodes is drawn if they interact. The basis of this approach, rooted in statistical physics, relies on the simplicity of the model to predict the behaviour of a system as a whole from the properties of its constituents. One of the main difficulties is that there must be no ambiguity on the individuals and their interactions definitions.

This basic model can however be more sophisticated. In practice, there may be more than one different type of nodes or more than one different type of links in a network. Both may have a variety of properties associated with them. They can carry weights. Links can be directed, i.e., pointing in one direction. They can join more than two nodes together. In such graphs called hypergraphs, the edges are referred as hyper edges. Graphs may also contain nodes of distinct types. For example, bipartite graphs contain two types of node with links running only from one type to the other. Affiliation networks in which people are joined together by common membership of groups can be represented using such a graph. The two types of nodes represent the people and the groups and a link account for the fact that an individual belong to one group. And there are many other levels of sophistication one can add. We will see examples of at least some of the variations described here in this issue.

Traditionally, networks have been studied extensively in the social sciences. Typically, networks in which nodes represent individuals and edges the interactions between them are build based on questionnaire where respondent detail their interactions with others. However, driven by the availability of computers and communication networks that allow us to gather and analyse data on a very large-scale, the focus is shifting away from the analysis of the properties of individual vertices or edges in small graphs to consideration on statistical properties of networks with millions or even billions of nodes.

Fuelled by the data explosion that we witness today, from social media to cell biology, the network-based paradigm is reshaping our approach to complexity. It aims at understanding the origins and characteristics of networks describing various complex systems. Based on data from the World Wide Web, activity on human gene and protein, mobile-phone records, import export data, stock data, social networks, software design, diseases spreading it led to the discovery that despite their many differences most complex systems are governed by common laws that determine their behaviour.

Many concepts and measures have been proposed to capture in quantitative terms their underlying organising principles. We briefly recall the most prominent concepts that characterise structure and behaviour of most real-world complex systems:

- *Low separation degree*: Despite their often large size, in most networks there is a relatively short path between any two nodes. The so called ‘small world’ property dates to the work the social psychologist Stanley Milgram. In an experiments participant of an acquaintance network were asked to pass a letter to one of their direct acquaintances in an attempt to get it to an assigned target individual. Most of the letters in the experiment were lost, but about a quarter reached the target and passed on average through the hands of only about six people. Well known as the ‘six degrees of separation’ concept the small-world property appears to characterise most complex networks: In the internet a computer can be reached on average through six routers, the co-authors in mathematics are on average within four authors from each other.
- *Heterogeneous distribution of node’s degree*: The *degree* of a node refers to the number of links attached to this node. It can be apprehended as a measure of the leadership of a node in the network. Highly connected nodes referred as hubs or authorities being critical elements of the networks. The *degree distribution* of a network $P(k)$, which gives the probability that a randomly selected node has exactly k edges, can be estimated by an histogram of the degrees of nodes. One of the most interesting developments in our understanding of complex networks is that they exhibit an inhomogeneous distribution with few nodes highly connected and the great majority of nodes poorly connected. In particular, for a large number of networks, it can be adequately described by a *power-law distribution*. Such networks are often referred as ‘scale free networks’. The ‘preferential attachment principle’ allows to explain the appearance of the power law distribution. It refers to the idea that if a node has a high degree, it has a higher probability to attract more connections and thus its connectivity grows at a faster rate than other nodes with low connectivity. Another common term used to refer to this principle is ‘the rich get richer’.
- *Assortativity*: While the degree of a node measure the number of links connected to it, the degree correlation measures the tendency of nodes to mix together according to their degree value. A network is said *assortative* when nodes relates principally with nodes with the same characteristics. It is *disassortative* when they tend to attach to others with different characteristics. Social networks appear to be assortative while other types of networks (information networks, technological networks, biological networks) appear to be disassortative.
- *High clustering coefficient*: it is usually observed that in a social network of acquaintance that ‘the friend of your friend is likely to be your friend’. This inherent tendency to cluster in circles of friends in which every member knows every other member is quantified by the *clustering (transitivity) coefficient*. This concept known also as ‘fraction of transitive triples’ in sociology is used to capture the degree of social embeddedness that characterises the nodes of a network.
- *Low density*: the density of a graph is given by the ratio of the number of existing links to the number of possible links in the graph. In the case of a simple undirected graph composed of N nodes, the number of possible links is $N*(N-1)/2$. In most observed real complex networks, graphs are very sparse having a number of links proportional to N .

In addition to these basic features, it is widely assumed that most networks exhibit a mesoscopic level of organisation, called ‘modules’ or ‘communities’. Indeed, communities have always been ubiquitous as elementary forms of organisation both in society and nature. Indeed, society offers a wide variety of possible group organisations. In biology, group of proteins having the same specific function within the cell can be identified in protein-protein interaction networks. In the World Wide Web, communities correspond to groups of pages dealing with the same or related topics and so on. Exploring network community structure is crucial to the understanding of the structural and functional properties of networks. It may help to formulate realistic mechanisms for its genesis and evolution and to better understand dynamic processes taking place on the network.

Although there is a tremendous effort for the inference of community structure from various scientific communities, there is no formal consensus on a definition capturing the gist of communities. In most case it is usually based on the basic idea that members of such a set form a cohesive group interacting with each other more ‘intensely’ than with those outside the group. As there is many and diverse understandings on how cohesiveness translates in formal graph-theoretic terms, many algorithms have appeared in the literature to discover networks community structure. One comprehensive state of the art of main approaches for community detection is presented in Fortunato (2010). Most of existing work is focused on partitioning a network into disjoint communities in simple, static large-scale graphs. In disjoint communities, a node belongs to only one group. However, in some real networks a given node might belong to different communities. For example, a person might simultaneously belong to many different communities. In biology, a large fraction of proteins belong to several protein complexes simultaneously. In order to reflect more precisely this intuition, there is a growing interest in the study of overlapping communities in the recent years.

Besides the distinction between overlapping and non-overlapping communities most contributions consider networks with unlabelled nodes and edges (edges can be weighted or not). However, in many real-world settings, networks can be composed of different types of nodes and edges, as it is the case in bibliographical networks (Kanawati, 2012). They can also be multi-partite as it is the case in folksonomies and social resource sharing, such as Flickr, Delicious, or CiteULike, where interactions are represented by tripartite graphs linking users to resource and tags used to annotate resources by hyperlinks.

Apart from these aspects, a great body of work focuses on the idea of finding a particular community structure relying on global knowledge of the entire network’s structure. This constraint can be problematic for networks which are too large and evolve too quickly to have a fully known structure. Indeed, nowadays, huge sets of data are accumulating at a tremendous pace in various fields of human activity. The typical size of large networks such as social network services, mobile phone networks or the web now counts in millions when not billions of nodes. In order to overcome these limitations, local methods attempt to find community inside the neighbourhood of a given node. For example, in a social network, these ‘user centric’ methods aim to quantify the local communities of a person.

The last years have witnessed a huge amount of research mainly focused on community structure discovery. Many techniques and quality criteria have been proposed. However, little work has been done to characterise the main properties of

real-world network community structure and to create models of such networks. This aspect of the problem that can help us to understand the meaning of these properties and how they interact is of prime interest.

Traditionally large networks with no apparent design principles have been described as random graphs. However, there are clear indications that real-world networks evolve following self-organising principles and evolutionary laws that cross-disciplinary boundaries. These organising principles are at some level encoded in their topology. Even if random graphs studied extensively by Erdős and Rényi exhibit the small world property, other topological properties is far those measured in real-world networks. First of all, they do not exhibit a community structure. Furthermore, their degree distribution is homogeneous while complex networks exhibit an inhomogeneous distribution. Such heterogeneity is responsible for a number of remarkable features of real networks, such as resilience to random failures and epidemic spreading. Therefore, it is important for the relevance of the simulation results, as well as the theoretical ones, to use realistic topologies.

Many models of complex networks have been proposed in the literature to overcome the random model limitations. Some of them attempt to explain the general properties we have cited. Others are aimed at reproducing a specific network of special interest. In general, it appears that there is no single model which is able to simultaneously capture all topological properties of complex networks. One of the most influential models is the so called Barabasi-Albert model (Barabasi and Albert, 1999). It exploits two mechanisms of network dynamics which captures the scale-free structure of the networks. First, most networks grow through the addition of new nodes, that link to nodes already present in the system. Second, according to the preferential attachment principle, more popular vertices of a network attract more new vertices.

Building on these models, algorithms that account for the heterogeneity of node degree distribution and community size that are observed in real networks have been proposed. Solutions for directed, weighted and overlapping community structured networks have been also been investigated. However, as the focus has been mainly on community structure discovery much more work need to be done in order to better understand the topological properties of communities and their effect on systems behaviour. Such developments are crucial for the design of models that can produce more realistic community structured networks.

The articles in this special issue reflect very well the main concerns on the subject: community structure discovery, modelling and characterisation of the community structure of real-world networks.

The first two papers address the problem of community detection in two different angles. While the first paper focuses on local community structure, the second offers a solution to unfold the overall hierarchical structure of overlapping communities.

In the article ‘Towards multi-ego-centred communities: a node similarity approach’ Maximilien Danisch, Jean-Loup Guillaume and Bénédicte Le Grand introduce a community detection algorithm focused on the groups related to one node (*ego-centred community*). Instead of using a cost function approach, as usually done, they relate the community structure to irregularities in the similarity between the node of interest and all the other nodes. In order to deal with large-scale networks their algorithm relies on an efficient similarity metric called the *carryover opinion metric*. This work naturally leads to the starting point of a promising new vision of communities: The

multi-ego-communities which extend the concept of ego-communities to the communities shared by a small set of nodes.

In their paper entitled ‘Multi-scale community detection using stability optimisation’, Erwan Le Martelot and Chris Hankin investigate an alternative to the classical modularity quality measure used to assess the quality of a partition in order to uncover the multiple level of organisation and the overlapping community structure of real-world networks. One of the advantages of stability is that it encompasses modularity and can be optimised similarly. In a greedy optimisation context, they demonstrate the benefit of using stability in order to clearly identify the community structure of relevant networks under test. Given the ubiquity of modularity in the community detection literature, this work may give rise to numerous extensions.

The third and fourth papers deal with community structured network modelling. In the third paper, global models for non-overlapping community structured networks are investigated in the context community detection algorithms evaluation. The fourth paper present a growing hypergraph model for affiliation networks with overlapping communities.

In ‘Towards realistic artificial benchmark for community detection algorithms evaluation’, Günce Keziban Orman, Vincent Labatut and Hocine Cherifi tackle the problem of community detection algorithm evaluation through the effectiveness of generative models capturing essential properties of real-world networks. They investigate the impact of the realism level of the network topological properties on the efficiency of non-overlapping community detection algorithms. This study provides a better understanding on which detection algorithm to use in a given real-world case.

The paper ‘Characterising and modelling social networks with overlapping communities’ of Dajie Liu, Norbert Blenn and Piet Van Mieghem propose a generative model dedicated to affiliation networks. Based on the hypergraph representation, the model uses the growing and preferential attachment mechanism in order to generate artificial networks that reproduce the properties of typical real-world affiliation networks such as co-authorship networks, actor collaboration networks and software collaboration networks. This work fills gaps of previous solutions by a more appropriate characterisation of networks topological properties.

The last two papers are dedicated to the investigation of real-world networks structure. Interaction networks representing the composition process of web services are analysed in the fifth paper, while the multidimensional nature of news clips is explored on the last one.

In ‘Community structure in interaction web service networks’, Chantal Cherifi and Jean-François Santucci introduce two web services network representation in order to understand the topological landscape of both syntactic and semantic web service coupled through the web in order to create new value-added services. Experimental results indicate that these networks exhibit the non-trivial topological properties of most real-world complex systems. The community structure revealed is a promising feature. It can be used to develop more efficient strategies in the publication and discovery processes of the continuously growing web services space.

In their paper entitled ‘The community structure of a multidimensional network of news clips’, José Luís Devezas and Álvaro Reis Figueira explore the multiple relation types that can occur between two entities. In order to study the impact of these different relationships on the community structure, they analysed a network of news clips where

the nodes are the news clips and a link is drawn when two nodes refer to either the same place, or the same people or the same date. The comparative evaluation of the community structure of the multidimensional approach where each of the three independent edge dimension are properly fused with the mono-dimensional one demonstrate the advantage of the former in order to identify semantically meaningful communities. This study highlights the value of considering the multiple aspects of relationship in order to better apprehend the community structure.

The web is no longer a simple documentary system. It grew to become a huge network of distributed data, applications and users. The outburst of social functionalities in web-based applications has fostered the deployment of shared spaces where people freely gather and interact with each other. However, this ever growing amount of information is a major challenge to achieve an efficient use of web technologies. This very complex system requires a multidisciplinary scientific approach. To handle this complexity a unifying network-based paradigm that is as useful in search engines as in cell biology is emerging. Coming from five different countries (France, UK, Turkey, The Netherlands and Portugal) 17 authors share their research with you and contribute to the primary focus of this special issue: to raise awareness of the web-based communities to the various analytical tools and algorithms that network theory is building in order to understand the community structure of real-world complex networks. We are convinced that it will help to a better understanding of many complex systems arising in various disciplines of human activity.

References

- Barabasi, A. and Albert, R. (1999) 'Emergence of scaling in random networks', *Science*, Vol. 286, No. 5439, p.509.
- Bialik, C. (2007) 'Sorry, you may have gone over your limit of network friends', *The Wall Street Journal Online*, 16 November.
- Dunbar, R.I.M. (1993) 'Coevolution of neocortical size, group size and language in humans', *Behavioral and Brain Sciences*, Vol. 16, No. 4, pp.681–735.
- Fortunato, S. (2010) 'Community detection in graphs', *Physics Reports*, Vol. 486, Nos. 3–5, pp.75–174.
- Kanawati, R. (2012) 'Mining the dynamics of scientific publication networks for collaboration recommendation', *Second International Workshop on Mining Communities and People Recommender (COMMPER@ ECML2012)*, 28 September, Bristol.
- Milgram, S. (1967) 'The small world problem', *Psychology Today*, Vol. 2, No. 1, pp.60–67.
- Russell, B.H., Shelley, G.A. and Killworth, P. (1987) 'How Much of a network does the GSS and RSW dredge up?', *Social Networks*, Vol. 9, No. 1, pp.49–63.
- Sipser, M. (1997) *Introduction to the Theory of Computation*, pp.248–271, Sections 7.4–7.5 (NP-completeness, additional NP-complete problems), PWS Publishing, Boston, ISBN: 0-534-94728-X.
- Snijders, T.A.B., Pattison, P.E., Robins, G. and Handcock, M.S. (2006) 'New specifications for exponential random graph models', *Sociological Methodology*, Vol. 36, No. 1, pp.99–153.